Lung Cancer Prediction by Using Machine Learning Models with Distributed System and Weka Visualization

Şenol PARUĞ[1]

, Fathia A. H. LAZRAG[2]

, Abdulhamid JABRR [2]

, Ebtisam M.H. FAKROUN[3]

[1] Kastamonu University, Science Faculty, Biology Department, Kastamonu/Türkiye

[2] Kastamonu University, Institute of Science and Technology, Department of Aquaculture, Kastamonu/Türkiye

[3] Misurata University, Ebtisam Mohamed Fakroun, Department of Engineering and Information Technology

*E-mail: tdebbek@yahoo.com

ABSTRACT

Lung cancer is one of the most common and deadly types of cancer worldwide. Modern lifestyle, carcinogens, smoking, and air pollution are the biggest causes of lung cancer. Early detection and accurate diagnosis are essential for effective treatment and improved survival rates. Machine learning models have shown great potential in predicting lung cancer, which can help in early detection and personalized treatment planning. This technology uses algorithms to analyze large amounts of data, including medical images, patient history, and genetic information to identify patterns and predict outcomes. In this paper, we will explore the use of machine learning models for lung cancer prediction, based on distributed systems and the Python language. In this protection, an implement on Predicting lung cancer using machine learning techniques based on distributed system application, using Python language

Keyword: - Lung Cancer Prediction, Machine Learning, Models, Distributed System, Weka Visualization

1. Introduction

Lung cancer is the uncontrolled growth of abnormal cells in one or both lungs (Kaur and Garg, 2021); (Rustam and Kharis, 2020); (Jagdale Swati et al., 2023); (Wang et al., 2022); (Hadisaputri et al., 2020). These abnormal cells do not perform the functions of normal lung cells and do not develop into healthy lung tissue. Tumors form and impede lung function (Kaur and Garg, 2021); (Rustam and Kharis, 2020). **Furthermore,** smoking Cigarettes are the main cause of lung cancer, accounting for about 85% of lung cancer cases. The risk of developing lung cancer varies according to the number of cigarettes smoked and the number of years of smoking (Wang et al., 2022); (Hadisaputri et al., 2020). However, some heavy smokers do not develop lung cancer. The risk of lung cancer decreases in people who quit smoking, but ex-smokers will continue to have a higher risk of developing lung cancer than people who have never smoked (Kaur and Garg, 2021); (Rustam and Kharis, 2020); (Jagdale Swati et al., 2023); (Hadisaputri et al., 2020).

About 15-20% of people who develop lung cancer have never smoked or smoked lightly. It is not known why these people develop lung cancer, but certain genetic mutations may be responsible. Lung cancer is a type of cancer that begins in the cells of the lungs, the vital organs responsible for breathing (Kaur and Garg, 2021); (Rustam and Kharis, 2020); (Jagdale Swati et al., 2023). It is one of the most prevalent and

deadliest forms of cancer worldwide. Lung cancer can develop in both smokers and non-smokers, although it is strongly associated with tobacco smoke (Hadisaputri et al., 2020). There are two main types of lung cancer: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (Che et al., 2020); (Zadian et al., 2021) . Non-small cell lung cancer is the most common type, accounting for approximately 80-85% of all cases. It usually grows and spreads more slowly than small-cell lung cancer. Small cell lung cancer, on the other hand, is less common and tends to grow and spread rapidly (Che et al., 2020); (Zadian et al., 2021) .

The primary cause of lung cancer is cigarette smoking, including both active smoking (smoking cigarettes directly) and passive smoking (inhaling secondhand smoke). Other risk factors include exposure to certain chemicals and substances like asbestos, radon gas, arsenic, and diesel exhaust (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020). Additionally, a family history of lung cancer, previous radiation therapy to the chest, and certain genetic mutations can also increase the risk. Lung cancer may not exhibit obvious symptoms in its early stages, which makes it difficult to detect and diagnose. However, as the disease progresses, common symptoms may include persistent cough, coughing up blood, chest pain, shortness of breath, fatigue, unexplained weight loss, and recurrent respiratory infections (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020). If any of these symptoms persist, it is crucial to consult a healthcare professional for further evaluation. Diagnosis of lung cancer typically involves a combination of imaging tests such as chest X-rays, CT scans, and MRIs (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020), as well as biopsies to examine the lung tissue for cancer cells. Once diagnosed, the stage of the cancer is determined to guide treatment decisions. Lung cancer is staged based on the size and location of the tumor, the presence of lymph node involvement, and whether it has spread to other parts of the body (Liu et al., 2020); (Bhateja et al., 2019). Treatment options for lung cancer depend on the type and stage of the disease. They may include surgery, radiation therapy, chemotherapy, targeted therapy, immunotherapy, or a combination of these approaches. The choice of treatment aims to remove or destroy cancer cells, alleviate symptoms, and improve overall survival rates. Palliative care is also provided to manage symptoms and improve the quality of life for patients with advanced or metastatic lung cancer (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020).

Prevention plays a crucial role in reducing the risk of developing lung cancer. Avoiding tobacco smoke, including quitting smoking and staying away from secondhand smoke, is the most effective preventive measure (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020). Additionally, minimizing exposure to environmental and occupational carcinogens, such as radon and asbestos, can also help reduce the risk. Early detection through screening programs can improve survival rates for individuals at high risk, such as long-term smokers (Liu et al., 2020); (Bhateja et al., 2019); (Chen et al., 2020); (Sehgal et al., 2020). Screening methods include low-dose computed tomography (LDCT) scans, which can detect lung cancer at an earlier stage when it is more treatable. Lung cancer is a serious and potentially life-threatening disease that affects the lungs. It can be caused by various factors, with tobacco smoke being the primary culprit. Early detection, timely diagnosis, and appropriate treatment are essential in improving outcomes for individuals affected by lung cancer (Liu et al., 2020); (Bhateja et al., 2019).

1.1 SYSTEM ARCHITECTURE

The lung cancer prediction system consists of machine learning algorithms and a data set, which the system trains in order to form a model (Raoof et al., 2020); (Chaturvedi et al., 2021); (Cui et al., 2020); (Riquelme and Akhloufi, 2020); (Shanthi and Rajkumar, 2021). The model is a program or algorithm that utilizes a set of data that enables it to recognize certain patterns (Raoof et al., 2020); (Chaturvedi et al., 2021). This allows it to reach a conclusion or make a prediction when provided with sufficient information, often a huge amount of data (Cui et al., 2020); (Riquelme and Akhloufi, 2020); (Shanthi and Rajkumar, 2021). In this paper, these processes will be used using distributed systems using the Ray framework.

1.2 Dataset:

The data set, Lung Cancer Prediction, was formed by collecting and organizing previous data, for people with lung cancer, or for people who were suspected of having lung cancer, (Raoof et al., 2020); (Chaturvedi et al., 2021); (Cui et al., 2020); (Riquelme and Akhloufi, 2020); (Shanthi and Rajkumar, 2021),

of different ages, and this data is of high value, as it helps in predicting the incidence of lung cancer (Raoof et al., 2020); (Chaturvedi et al., 2021). These data were entered into the machine learning algorithm, the programs were trained on previous models, and knowledge was extracted from it to give the correct prediction, for the new data that was not entered. in the system before, this data was stored in a database, where it was collected, and turned into a data file, so that the machine learning algorithm could deal with it, Dataset as shown in Figure (1). below.

Name	Surname	Age	Smokes	AreaQ	Alkhol	Result
John	Wick	35	3	5	4	1
John	Constanti	27	20	2	5	1
Camela	Anderson	30	0	5	2	0
Alex	Telles	28	0	8	1	0
Diego	Maradona	68	4	5	6	1
Cristiano	Ronaldo	34	0	10	0	0
Mihail	Tal	58	15	10	0	0
Kathy	Bates	22	12	5	2	0
Nicole	Kidman	45	2	6	0	0
Ray	Milland	52	18	4	5	1
Fredric	March	33	4	8	0	0
Yul	Brynner	18	10	6	3	0
Joan	Crawford	25	2	5	1	0
Jane	Wyman	28	20	2	8	1
Anna	Magnani	34	25	4	8	1
Katharine	Hepburn	39	18	8	1	0
Katharine	Hepburn	42	22	3	5	1
Barbra	Streisand	19	12	8	0	0
Maggie	Smith	62	5	4	3	1
Glenda	Jackson	73	10	7	6	1
Jane	Fonda	55	15	1	3	1
Maximiliar	Schell	33	8	8	1	0
Gregory	Peck	22	20	6	2	0
Sidney	Poitier	44	5	8	1	0
Rex	Harrison	77	3	2	6	1
Lee	Marvin	21	20	5	3	0
Paul	Scofield	37	15	6	2	0

Figure (1) Screenshot of MY-SQL file as a data set of lung cancer.

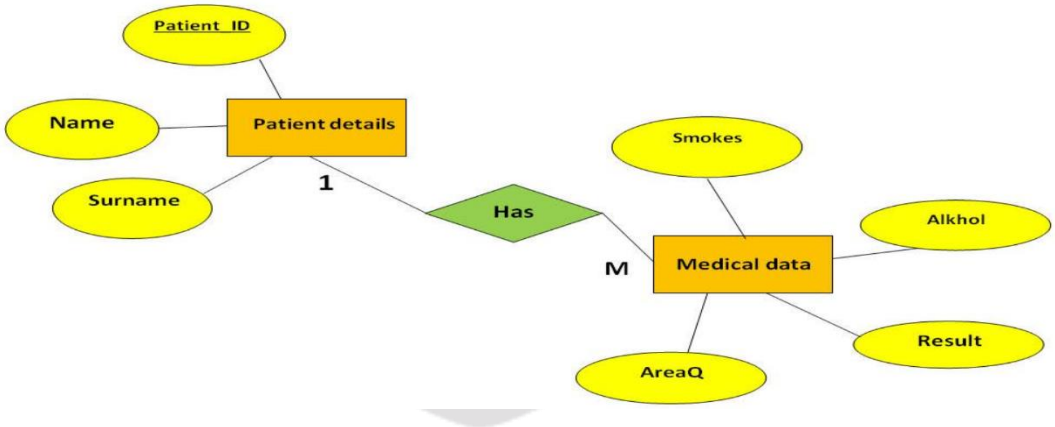

Figure. (2). Entity relationship diagram (ERD) of the lung cancer files.

2. RAY framework

RAY is an open source project of a language Python for parallel programming and distributed systems (Aziz et al., 2021); (Kim et al., 2020); (Sahin, 2019). It is one of the modern systems and frameworks that helps to prepare software based on the logic of working parallelism and distributed systems (Aziz et al., 2021); (Kim et al., 2020); (Sahin, 2019), for which parallel and distributed processing is an integral part of modern applications (Aziz et al., 2021); (Kim et al., 2020). This research needs to use multiple cores or multiple machines to accelerate applications or run them at scale (Aziz et al., 2021); (Kim et al., 2020). The software and network application infrastructure and query and query-response systems are not single-

threaded programs running on someone's laptop, but rather a set of services that communicate and interact with one another. distributed and running in parallel. Shown in the Figure (3) below.

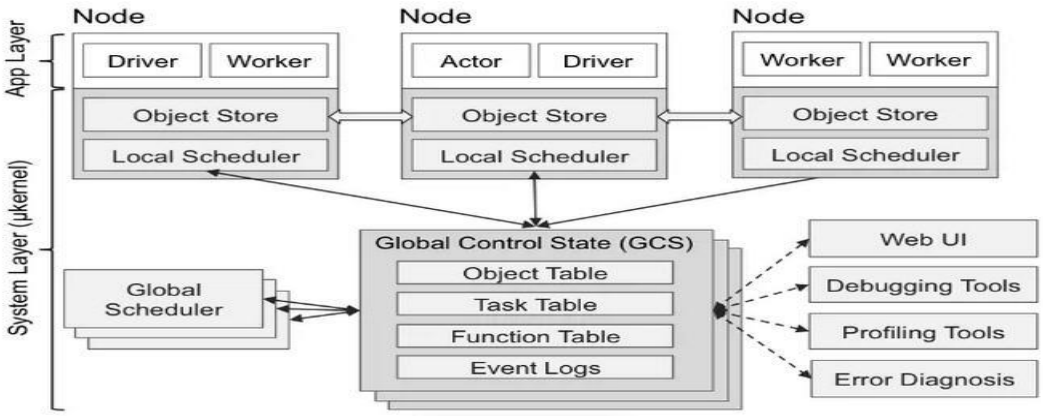

Figure (3) RAY Framework Architecture (https://www.mecbot.ai/scaling-python-modules-using-ray-framework/).

In this project, the RAY framework will be used, and defined (locally within the computer), (Aziz et al., 2021); (Kim et al., 2020) because the RAY framework is a virtual environment containing 4 cores, each core representing a remote computer.

2.1 Machine Learning Concepts

Machine learning, one of the branches of artificial intelligence, is constantly evolving (Abdulkareem et al., 2019); (Janiesch et al., 2021); (Doupe et al., 2019); (Hüllermeier and Waegeman, 2021); (Alam, 2022). Machine learning helps a person make the right decision, especially in cases of cancer diagnosis, where machine learning can be used for a greater understanding of real-world events (Abdulkareem et al., 2019); (Janiesch et al., 2021); (Doupe et al., 2019); (Hüllermeier and Waegeman, 2021). For example, suppose a person comes to the doctor and reports that they have lung cancer. The doctor based on his belief system which he has learned using his experience and knowledge predicts (essentially decides) whether or not a person will have lung cancer (Abdulkareem et al., 2019); (Janiesch et al., 2021); (Doupe et al., 2019); (Hüllermeier and Waegeman, 2021); (Alam, 2022). Depending on the clinician's expertise, to increase the reliability of the assessment, we can then replace the "human belief system" with the AI/ML system (one or more models) and the "experience and knowledge" with the data fed into this AI/ML system (Doupe et al., 2019); (Hüllermeier and Waegeman, 2021); (Alam, 2022). Doctors can also use ML models trained using historical data along with their experience and intelligence to predict whether or not a person will develop a disease. When human and machine learning intelligence are used together, this is why it is also called an augmentative system (Abdulkareem et al., 2019); (Janiesch et al., 2021); (Doupe et al., 2019). The doctor's decision to diagnose lung cancer. Machine learning differs from traditional programming, in the way the output is, in the case of machine learning it is the rules that are based on predictions, through training and testing. For this, input and output data are used with machine learning algorithms to output (model). Figure (4) below.

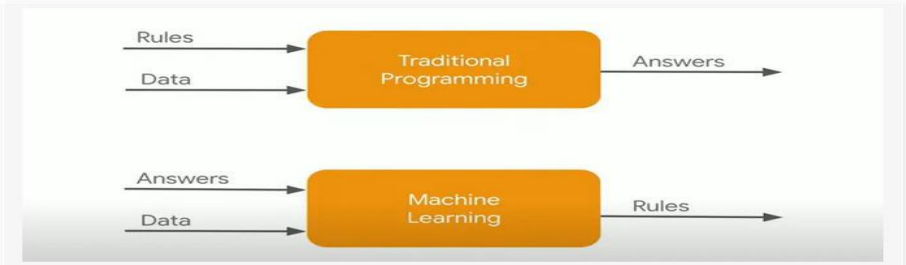

Figure (4) Difference between Machine learning and traditional programming

3. Types of machine learning

Machine learning involves showing a large amount of data to a machine so that it can learn and predict, find patterns, or classify the data (Alloghani et al., 2020); (Haidari et al., 2021); (Ajay et al., 2022); (Dhall et al., 2020); (Palacio-Niño and Berzal, 2019); (Rozemberczki et al., 2020). The algorithm used determines the type of machine learning, which works a little differently. There are two types of machine learning: supervised learning and unsupervised learning (Alloghani et al., 2020); (Heidari et al., 2021); (Ajay et al., 2022); (Dhall et al., 2020).

3.1 Supervised learning

This type of machine learning gets its name because the machine is "supervised" while learning, which means that the research feed information to the algorithm to help it learn (Alloghani et al., 2020); (Heidari et al., 2021); (Ajay et al., 2022); (Dhall et al., 2020); (Palacio-Niño and Berzal, 2019); (Rozemberczki et al., 2020). The score you provide to the device is compiled on the data, and the rest of the information you provide is used as input features. Supervised learning is effective for a variety of business purposes, including sales forecasting, inventory optimization, and fraud detection, and medical purposes such as lung cancer prediction (Alloghani et al., 2020); (Heidari et al., 2021).

3.2 Unsupervised learning

While supervised learning requires users to help the machine learn, unsupervised learning does not use the same training sets and labeled data. Instead, the machine looks for less obvious patterns in the data (Alloghani et al., 2020); (Heidari et al., 2021); (Ajay et al., 2022); (Dhall et al., 2020); (Palacio-Niño and Berzal, 2019); (Rozemberczki et al., 2020). This type of machine learning is very useful when the research need to identify patterns and use data to make decisions (Alloghani et al., 2020); (Heidari et al., 2021); (Ajay et al., 2022); (Dhall et al., 2020) Unsupervised learning can also compile the text document to extract the topic. This type of machine learning is widely used to create predictive models (Alloghani et al., 2020); (Heidari et al., 2021); (Ajay et al., 2022). Common applications also include aggregation, which creates a model that groups objects together based on certain properties, and association, which defines rules between collections, figure (4) shows types of machine learning (Palacio-Niño and Berzal, 2019); (Rozemberczki et al., 2020).

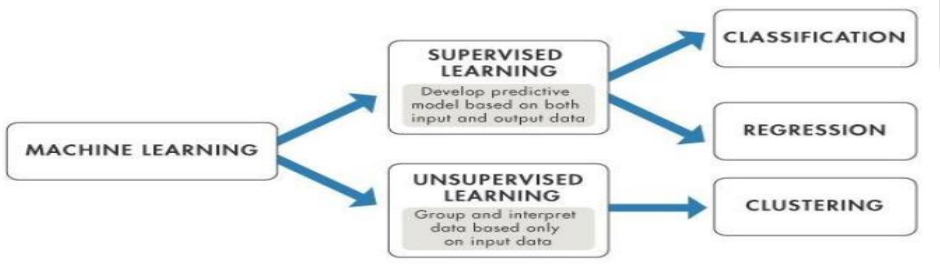

Figure (5) shows types of machine learning types adapted from (Alloghani et al., 2020); (Heidari et al., 2021).

The classification and Logistic Regression Classifier

This research idea into preset categories a.k.a "sub-populations." With the help of these pre-categorized training datasets, classification in machine learning programs leverage a wide range of algorithms to classify future datasets into respective and relevant categories (Shah et al., 2020); (Itoo and Singh, 2021); (Thabtah et al., 2019); (Robles-Velasco et al., 2020). Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories (Shah et al., 2020); (Itoo and Singh, 2021). The logistic

regression model is one of the most important statistical models for modeling the probability of a particular class or event such as success/fail (Shah et al., 2020); (Itoo and Singh, 2021); (Thabtah et al., 2019); (Robles-Velasco et al., 2020). This is because logistic regression uses several predicted variables which can be either continuous or categorical. This can be extended to model several classes of events such as determining if the image contains a cat, tiger, fish, etc. Each object detected in the image will be assigned a probability between 0 and 1, such that the sum total is equal to one. Logistic regression is also known by other names such as Logit model or general entropy classifier. Logistic regression falls under supervised machine learning algorithms intended for classification tasks. Below is an outline of the general use of logistic regression and other common linear classifiers (Shah et al., 2020); (Itoo and Singh, 2021); (Thabtah et al., 2019); (Robles-Velasco et al., 2020).

The Logistic Curve

The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ($\bar{Y}$). The formula to do so may be written either

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$ OR

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the natural logarithm (about 2.718) and a and b are the parameters of the model. The value of a yields P when X is zero, and b adjusts how quickly the probability changes with changing X a single unit (we can have standardized and un standardized b weights in logistic regression, just as in ordinary linear regression). Because the relation between X and P is nonlinear, b does not have a straightforward interpretation in this model as it does in ordinary linear regression (Shah et al., 2020); (Itoo and Singh, 2021).

Loss Function

A loss function is a measure of fit between a mathematical model of data and the actual data. We choose the parameters of our model to minimize the badness-of-fit or to maximize the goodness-of-fit of the model to the data. With least squares (the only loss function we have used thus far), we minimize SSres, the sum of squares residual. This also happens to maximize SSreg, the sum of squares due to regression. With linear or curvilinear models, there is a mathematical solution to the problem that will minimize the sum of squares, that is, b = **(X'** Or □ = $\mathbf{R^{-1}r}$

With some models, like the logistic curve, there is no mathematical solution that will produce least squares estimates of the parameters. For many of these models, the loss function chosen is called maximum likelihood. A likelihood is a conditional probability, for instance, (P(Y|X), the probability of Y given X). We can pick the parameters of the model (a and b of the logistic curve) at random or by trial-and-error and then compute the likelihood of the data given those parameters (actually, we do better than trail-and-error, but not perfectly). We will choose as our parameters, those that result in the greatest likelihood computed. The estimates are called maximum likelihood because the parameters are chosen to maximize the likelihood (conditional probability of the data given parameter estimates) of the sample data. The techniques actually employed to find the maximum likelihood estimates fall under the general label numerical analysis. There are several methods of numerical analysis, but they all follow a similar series of steps. First, the computer picks some initial estimates of the parameters. Then it will compute the likelihood of the data given these parameter estimates. Then it will improve the parameter estimates slightly and recalculate the likelihood of

the data. It will do this forever until we tell it to stop, which we usually do when the parameter estimates do not change much (usually a change .01 or .001 is small enough to tell the computer to stop). Sometimes we tell the computer to stop after a certain number of tries or iterations, for instance, 20 or 250. This usually indicates a problem in estimation. $\mathbf{X)^{-1}X'y}$

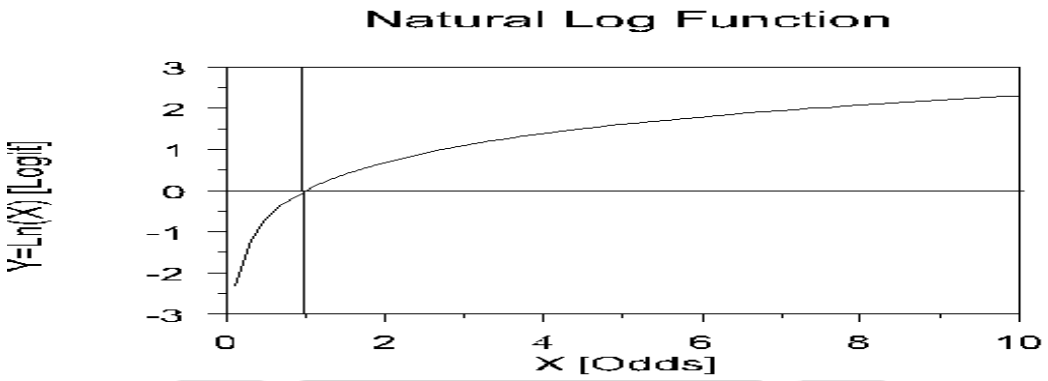

Figure. 6. explaining the logistic regression algorithm

After explaining the logistic regression algorithm, this algorithm was used to predict the incidence of lung cancer, in the data set, which was formed in hospitals, and from the data of patients with lung cancer. Now, this research shows a diagram to apply the logistic regression classifier algorithm in a project Predicting lung cancer using machine learning techniques depending on a distributed system application Figure (7).

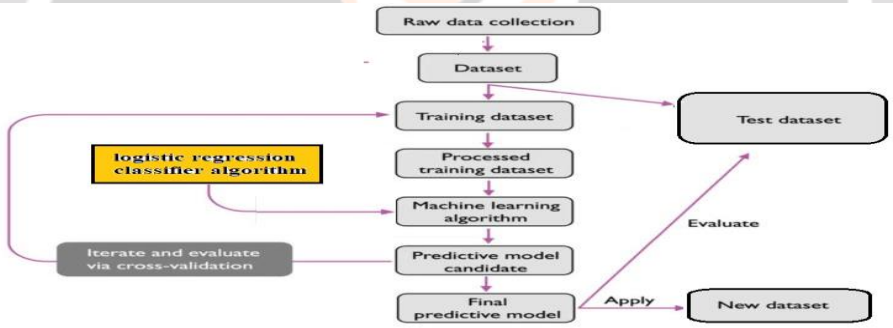

Figure (7) diagram to apply the logistic regression classifier algorithm in a project Predicting lung cancer using machine learning techniques.

GUI implements the system by Python language:

When the program is executed, the RAY framework is prepared to start operating, and 4 cores are prepared to work as separate computers. First, the first core fetches the data set from the source, and the second core prepares the pre-processing process for the data and then displays the data it retrieved. The third core is doing the training and testing group, for the logistic regression algorithm, and here the machine learning technique is used in Figure (8),

```
2023-05-10 09:01:48,597 INFO worker.py:1538 -- Started a local Ray instance.
(read_DataSet pid=12784) Dataset : (59, 7)
(read_DataSet pid=12784) <class 'pandas.core.frame.DataFrame'>
(read_DataSet pid=12784) RangeIndex: 59 entries, 0 to 58
(read_DataSet pid=12784) Data columns (total 7 columns):
(read_DataSet pid=12784)  #   Column   Non-Null Count  Dtype
(read_DataSet pid=12784) ---  ------   --------------  -----
(read_DataSet pid=12784)  0   Name     59 non-null     object
(read_DataSet pid=12784)  1   Surname  59 non-null     object
(read_DataSet pid=12784)  2   Age      59 non-null     int64
(read_DataSet pid=12784)  3   Smokes   59 non-null     int64
(read_DataSet pid=12784)  4   AreaQ    59 non-null     int64
(read_DataSet pid=12784)  5   Alkhol   59 non-null     int64
(read_DataSet pid=12784)  6   Result   59 non-null     int64
(read_DataSet pid=12784) dtypes: int64(5), object(2)
(read_DataSet pid=12784) memory usage: 3.4+ KB
(read_DataSet pid=12784) None
(read_DataSet pid=12784)         Name      Surname  Age  Smokes  AreaQ  Alkhol  Result
(read_DataSet pid=12784) 0       John         Wick   35       3      5       4       1
(read_DataSet pid=12784) 1       John  Constantine   27      20      2       5       1
(read_DataSet pid=12784) 2     Camela     Anderson   30       0      5       2       0
```

Figure (8) A screenshot of the pre-processing dataset and prepare mode in the background

Moreover, a model for logistic regression is prepared, and all these operations are worked in the background. After completing all the previous operations, the visual interface of Predicting lung cancer using machine learning techniques depending on a distributed system application Figure (9) below.

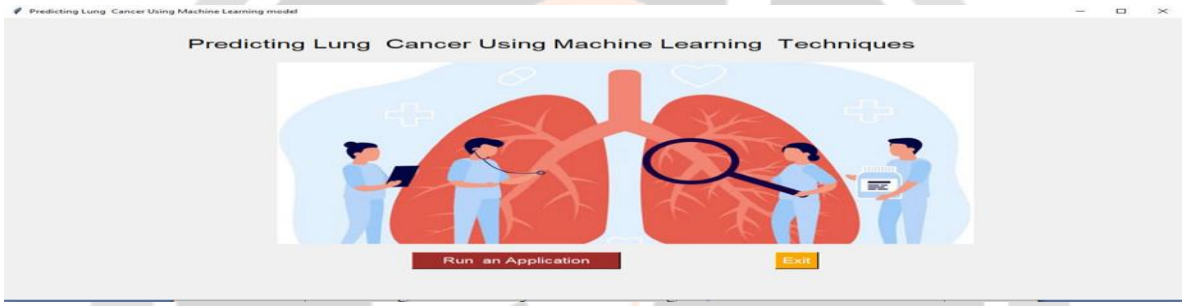

Figure (9) The screenshot of the GUI of Predicting lung cancer using Machine learning techniques

When the Run an Application button is pressed, the process of model for data by using logistic regression starts, where the fourth kernel in the Ray framework receives a model that has been trained and tested, performs the final evaluation process and shows the result of testing Score of the Model figure (10) below.

```
(create_model pid=16304) ===============Logistic Regression Model.....
(create_model pid=16304)
(create_model pid=16304) X train shape:  (53, 4)
(create_model pid=16304) Y train shape:  (53,)
(create_model pid=16304) X test shape:  (6, 4)
(create_model pid=16304) Y test shape:  (6,)
(create_model pid=16304) Logistic Regression Model Complete .....
(create_model pid=16304)
(create_model pid=16304) ================================================
(create_model pid=16304)
(create_model pid=16304) Testing  Score = 1.0
(create_model pid=16304) Ray ShutDown....
```

Figure (10) A screenshot logistic regression model Accuracy

After the implementation, the logistic regression algorithm is applied to the lung cancer dataset, and the number of samples between training and testing is clear. and tested the accuracy of the final model for predicting lung cancer.

4. EVALUATION

After completing the implementation of the program, and working on the lung cancer data set, a number of results emerged, the most important of which were: Figure (11) showing the number of cases of Infected and non-infected cases of lung cancer

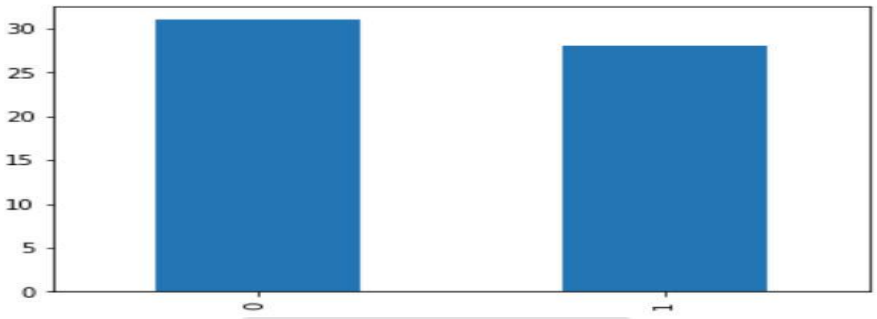

Figure (11) Infected and non-infected cases of lung cancer

In addition, the program shows charts of the most important attribute in the dataset of lung cancer, Figure (12) below.

Figure (12): an attribute view of lung cancer,

The logistic regression model predicts possible cases of lung cancer, and this is a practical application of one of the machine learning models in Python language and distributed systems.

Weka evaluation

Weka is a powerful data mining application that can help you better understand the data acquired (Verma, 2019); (Attwal and Dhiman, 2020); (Hammoodi et al., 2021). The application features powerful data analysis tools that can be used to extract information and develop new machine-learning schemes (Verma, 2019); (Attwal and Dhiman, 2020). Overall, the app is an excellent data analysis tool. Shown in Figure (13) below.

Figure (13): weka software

This example illustrates the use of k-means clustering with WEKA The sample data set used for this example is based on the "lung_canser_data" available in comma-separated format (lung_cancer_data.csv). This document assumes that appropriate data preprocessing has been performed. The resulting data file is " lung_cancer_data.csv " and includes 60 instances. As an illustration of performing clustering in WEKA, the research used its implementation of the K-means algorithm to cluster the data in this lung cancer data set, and to characterize the resulting customer segments (Verma, 2019); (Attwal and Dhiman, 2020); (Hammoodi et al., 2021). Figure (14) shows the main WEKA Explorer interface with the data file loaded.

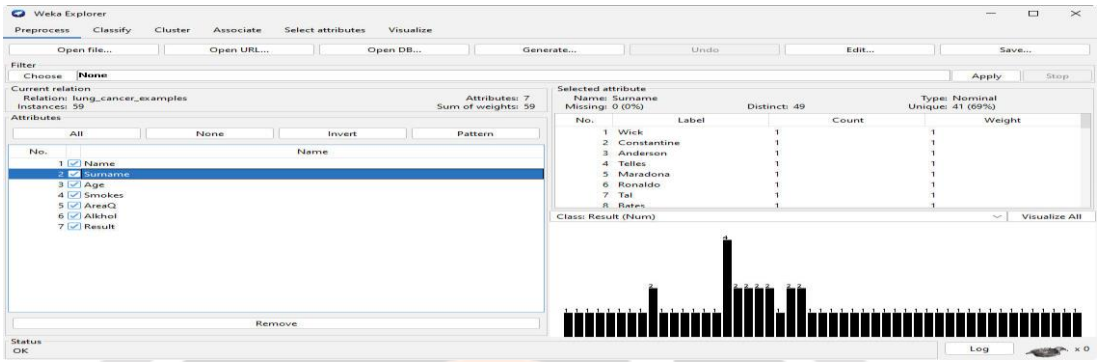

Figure (14) Main WEKA Explorer interface with the data file loaded.

All fields' in (lung_cancer_data.csv) file can be viewed by clicking on the visualize All button figure (15)

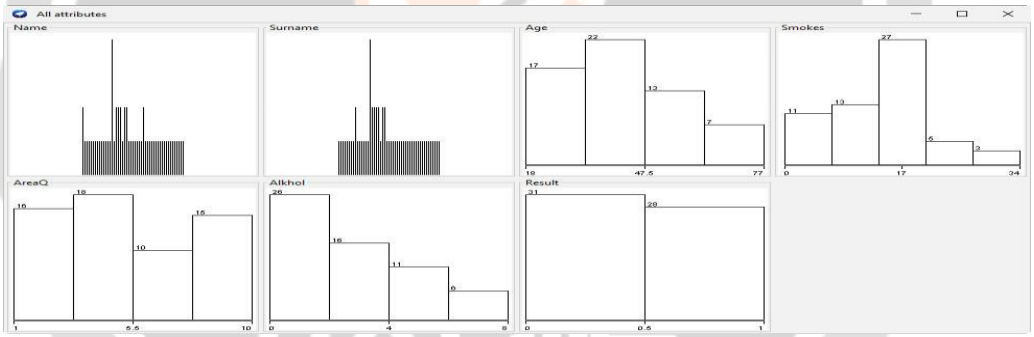

Figure (15) has visualize all Fields in select the "Cluster" tab, and select "Simple k-Mean". Figure (16) below.

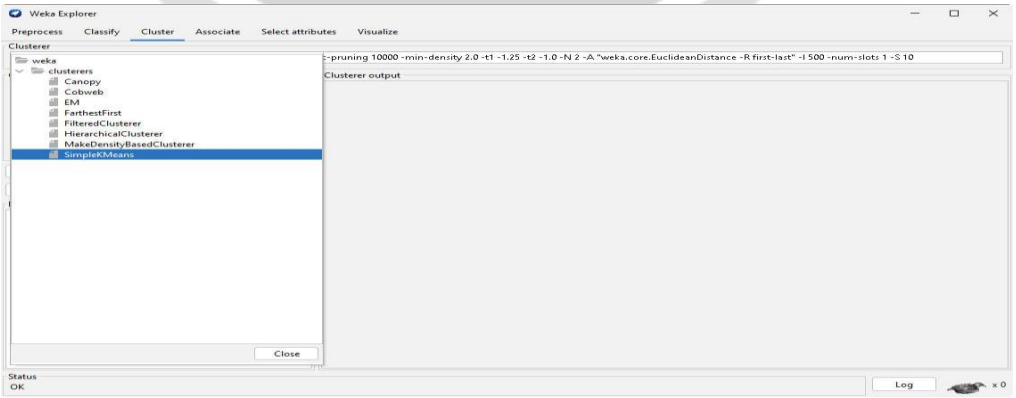

Figure (16) Select the "Simple k-Mean" algorithm. After that, the properties of the algorithm (simple k- mean) are chosen. The research has defined the number of clusters = 5, figure (17).

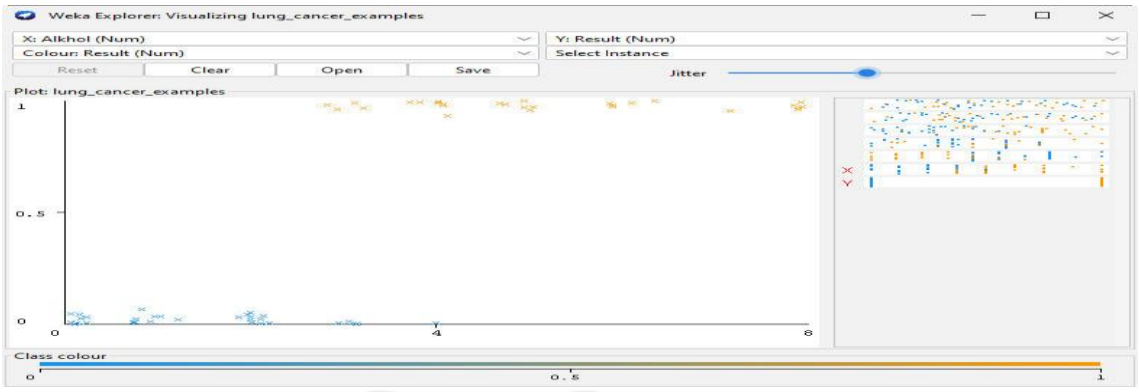

Figure (17) Simple k-Mean" algorithm cluster of data for the lung cancer dataset

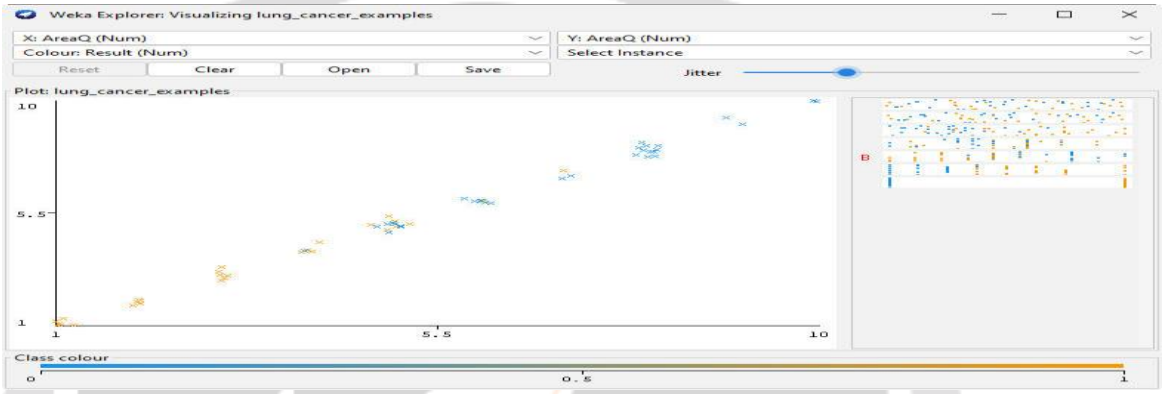

Figure (18) Screenshot of Canoby a logarithm cluster of data for the lung cancer dataset.

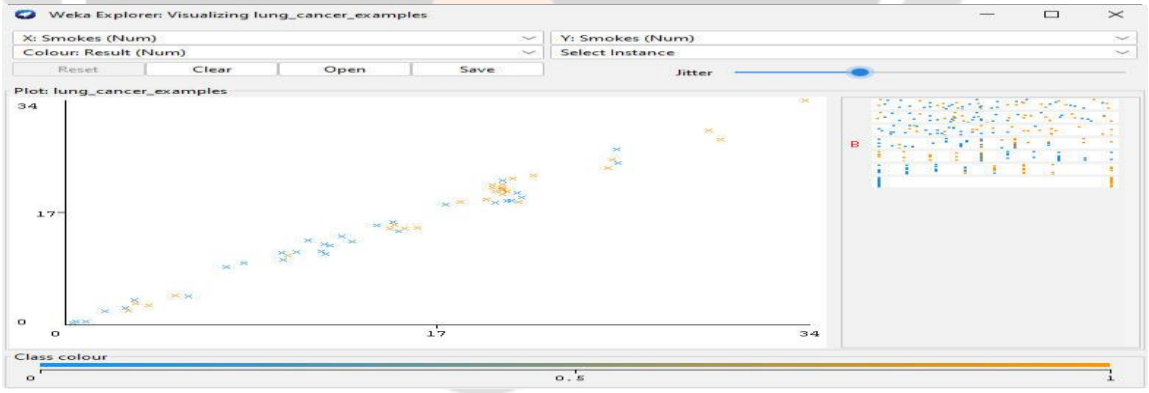

Figure (19) Screenshot of Hirercal cluster a logarithm cluster of data for the lung cancer dataset.

5. SCALABILITY AND RELIABILITY

After applying the lung cancer prediction system using machine learning techniques depending on distributed system application, and evaluating the results, it turned out that, working on parallel (distributed) systems using the beam framework helps in improving the speed of work, and this was shown through practical experience, whereas the process of loading the data set takes place in a remote program (a simulation of a separate computer), and the training and testing process of logistic regression on the data set takes place in a separate kernel, and the final evaluation process of the model takes place in a separate kernel, all these operations of Machine learning take place in the background, unnoticed by the system operator so that multi-core systems make optimal use of computer resources. One of the most important of these sources is the main

memory. Parallelism gives the advantage of speed in completing the work. Cloud computing can be used to speed up the completion of the work, which greatly helps in making calculations and evaluating the system quickly. This depends on the cost of the project and the budget allocated for expanding the system.

6. TRADE-OFF AND LIMITATIONS

In this project, a logistic regression classifier algorithm was implemented on a data set, specific to lung cancer, based on data collected from specialized treatment and diagnostic centers, where a local version of the RAY framework was used, as it provides a RAY framework with a maximum of 4 Centers to work on personal computers. Therefore, no more than 4 centers were made to implement the system, and the calculations showed the speed of implementation because it works in parallel, and this reinforces the hypothesis that distributed systems save time and effort, and increase the efficiency of results, by updating the workgroup data, through the branches for providing Data, and improving the training process, based on the data, will increase the accuracy of the final model, and this is one of the objectives of the project implementation.

7. FUTURE WORK

In the future, a lung cancer prediction system can be developed by machine learning, based on distributed systems, by using deep learning techniques. The Ray framework supports deep learning techniques, and this helps in creating new ideas, to increase the accuracy of the final model after the prediction process, and other algorithms can be used for the classification and comparison of results, to find out which techniques are better in terms of accuracy of the comparison result.

8. CONCLUSIONS

After completing the implementation of the Predicting lung cancer system using machine learning techniques depending on a distributed system application. , where four cores were used as far as the ray framework allows, in the local version, in the PC, where each core represents an independent remote machine, and it is noted by the program that, the process of training the lung cancer prediction model is done in the core, it is considered a remote machine remote, and the process of evaluating the model is done in another remote ready, and this realizes the idea of working in parallel with others. The importance of using computer resources, which is the advantage of memory management, increased performance, and speed of work completion, was taken advantage of with the application of the principle of distributed systems that work in parallel using the Python language.

9. REFERENCES

1. Abdulkareem, K. H., Mohammed, M. A., Gunasekaran, S. S., Al-Mhiqani, M. N., Mutlag, A. A., Mostafa, S. A., ... & Ibrahim, D. A. (2019). A review of fog computing and machine learning: concepts, applications, challenges, and open issues. Ieee Access, 7, 153123-153140.

2. Ajay, P., Nagaraj, B., Kumar, R. A., Huang, R., & Ananthi, P. (2022). Unsupervised hyperspectral microscopic image segmentation using deep embedded clustering algorithm. Scanning, 2022.

3. Alam, A. (2022, April). A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 69-74). IEEE.

4. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

5. Attwal, K. P. S., & Dhiman, A. S. (2020). Exploring data mining Tool-Weka and using Weka to build and evaluate predictive models. Advances and Applications in Mathematical Sciences, 19(6), 451-469.

6. Aziz, Z. A., Abdulqader, D. N., Sallow, A. B., & Omer, H. K. (2021). Python parallel processing and multiprocessing: A rivew. Academic Journal of Nawroz University, 10(3), 345-354.

7. Aziz, Z. A., Abdulqader, D. N., Sallow, A. B., & Omer, H. K. (2021). Python parallel processing and multiprocessing: A rivew. Academic Journal of Nawroz University, 10(3), 345-354.

8. Bhateja, P., Chiu, M., Wildey, G., Lipka, M. B., Fu, P., Yang, M. C. L., ... & Dowlati, A. (2019). Retinoblastoma mutation predicts poor outcomes in advanced non-small cell lung cancer. Cancer medicine, 8(4), 1459-1466.

9. Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021). Prediction and classification of lung cancer using machine learning techniques. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, No. 1, p. 012059). IOP Publishing.

10. Chen, B. T., Chen, Z., Ye, N., Mambetsariev, I., Fricke, J., Daniel, E., ... & Salgia, R. (2020). Differentiating peripherally-located small cell lung cancer from non-small cell lung cancer using a CT radiomic approach. Frontiers in Oncology, 10, 593.

11. Chen, B. T., Chen, Z., Ye, N., Mambetsariev, I., Fricke, J., Daniel, E., ... & Salgia, R. (2020). Differentiating peripherally-located small cell lung cancer from non-small cell lung cancer using a CT radiomic approach. Frontiers in Oncology, 10, 593.

12. Cui, L., Li, H., Hui, W., Chen, S., Yang, L., Kang, Y., ... & Feng, J. (2020). A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. BMC bioinformatics, 21, 1-14.

13. Dhall, D., Kaur, R., & Juneja, M. (2020). Machine learning: a review of the algorithms and its applications. Proceedings of ICRIC 2019: Recent Innovations in Computing, 47-63.

14. Doupe, P., Faghmous, J., & Basu, S. (2019). Machine learning for health services researchers. Value in Health, 22(7), 808-815.

15. Hadisaputri, Y. E., Cahyana, N., Muchtaridi, M., Lesmana, R., Rusdiana, T., Chaerunisa, A. Y., ... & Subarnas, A. (2020). Apoptosis-mediated antiproliferation of A549 lung cancer cells mediated by Eugenia aquea leaf compound 2', 4'-dihydroxy-6'-methoxy-3', 5'-dimethylchalcone and its molecular interaction with caspase receptor in molecular docking simulation. Oncology letters, 19(5), 3551-3557.

16. Hammoodi, M. S., Al Essa, H. A., & Hanon, W. A. (2021). The Waikato open source frameworks (WEKA and MOA) for machine learning techniques. In Journal of Physics: Conference Series (Vol. 1804, No. 1, p. 012133). IOP Publishing.

17. Heidari, M., Zad, S., & Rafatirad, S. (2021). Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.

18. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110, 457-506.

19. Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13, 1503-1511.

20. Jagdale Swati, C., HableAsawaree, A., & ChabukswarAnuruddha, R. (2023). Nanomedicine in lung cancer therapy. Advances in Novel Formulations for Drug Delivery, 433-448.

21. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.

22. Kaur, C., & Garg, U. (2021). Artificial intelligence techniques for cancer detection in medical image processing: A review. Materials Today: Proceedings.

23. Kim, T., Cha, Y., Shin, B., & Cha, B. (2020). Survey and performance test of python-based libraries for parallel processing. In The 9th International Conference on Smart Media and Applications (pp. 154-157).

24. Kim, T., Cha, Y., Shin, B., & Cha, B. (2020). Survey and performance test of python-based libraries for parallel processing. In The 9th International Conference on Smart Media and Applications (pp. 154-157).

25. Liu, S., Liu, S., Zhang, C., Yu, H., Liu, X., Hu, Y., ... & Fu, Q. (2020). Exploratory study of a CT radiomics model for the classification of small cell lung cancer and non-small-cell lung cancer. Frontiers in Oncology, 10, 1268.

26. Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. arXiv preprint arXiv:1905.05667.

27. Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020). Lung Cancer prediction using machine learning: A comprehensive approach. In 2020 2nd International Conference on innovative mechanisms for industry applications (ICIMIA) (pp. 108-115). IEEE.

28. Riquelme, D., & Akhloufi, M. A. (2020). Deep learning for lung cancer nodules detection and classification in CT scans. Ai, 1(1), 28-67.

29. Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliability Engineering & System Safety, 196, 106754.

30. Rozemberczki, B., Kiss, O., & Sarkar, R. (2020). Karate Club: an API oriented open-source python framework for unsupervised learning on graphs. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 3125-3132).

31. Rustam, Z., & Kharis, S. A. A. (2020). Comparison of support vector machine recursive feature elimination and kernel function as feature selection using support vector machine for lung cancer classification. In Journal of Physics: Conference Series (Vol. 1442, No. 1, p. 012027). IOP Publishing.

32. Sahin, F. E. (2019). Open-source optimization algorithms for optical design. Optik, 178, 1016-1022.

33. Sahin, F. E. (2019). Open-source optimization algorithms for optical design. Optik, 178, 1016-1022.

34. Sehgal, K., Varkaris, A., Viray, H., VanderLaan, P. A., Rangachari, D., & Costa, D. B. (2020). Small cell transformation of non-small cell lung cancer on immune checkpoint inhibitors: uncommon or under-recognized? Journal for immunotherapy of cancer, 8(1).

35. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5, 1-16.

36. Shanthi, S., & Rajkumar, N. (2021). Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods. Neural Processing Letters, 53, 2617-2630.

37. Supriya, M., & Deepa, A. J. (2020). Machine learning approach on healthcare big data: a review. Big Data and Information Analytics, 5(1), 58-75.

38. Thabtah, F., Abdelhamid, N., & Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. Health information science and systems, 7, 1-11.

39. Verma, A. (2019). Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. International Research Journal of Engineering and Technology, 5(13), 54-60.

40. Wang, T., Tang, T., Jiang, Y., He, T., Qi, L., Chang, H., ... & Wang, J. (2022). PRIM2 Promotes Cell Cycle and Tumor Progression in p53-Mutant Lung Cancer. Cancers, 14(14), 3370.

41. Zadian, S. S., Adcock, I. M., Salimi, B., & Mortaz, E. (2021). Circulating levels of monocytic myeloid-derived suppressor cells (M-MDSC) and CXCL-8 in non-small cell lung cancer (NSCLC). Tanaffos, 20(1), 15.